Volume 10 Issue 03 Published, August 10, 2006 ISSN 1535-864X DOI: 10.1535/itj.1003

Intel® Technology Journal

Intel® Virtualization Technology

New Client Virtualization Usage Models Using Intel® Virtualization Technology

More information, including current and past issues of Intel Technology Journal, can be found at:
http://developer.intel.com/technology/itj/index.htm

New Client Virtualization Usage Models Using Intel® Virtualization Technology

Mahendra Ramachandran, Digital Enterprise Group, Intel Corporation
Ned Smith, Digital Enterprise Group, Intel Corporation
Matthew Wood, Digital Home Group, Intel Corporation
Sharad Garg, Digital Home Group, Intel Corporation
Jim Stanley, Digital Home Group, Intel Corporation
Eswar Eduri, Software Solutions Group, Intel Corporation
Rinat Rappoport, Software Solutions Group, Intel Corporation
Arie Chobotaro, Software Solutions Group, Intel Corporation
Carl Klotz, Digital Enterprise Group, Intel Corporation
Lori Janz, Mobile Platforms Group, Intel Corporation

Index words: virtualization, operating systems, manageability, business computing, security

ABSTRACT

Intel's Embedded IT (EIT) strategy focuses on defining a set of usages aimed at benefiting IT departments and home PC customer support by providing advanced and remote capabilities for provisioning, manageability, diagnostics, and remediation of the client (desktop and mobile) platform. EIT leverages key platform technologies supported on Intel® vPro™ technology platforms and select digital home platforms, including Intel® Virtualization Technology$^{\Delta}$ (VT), Intel® Active Management Technology^{+} (AMT), and dual-core processors to deliver an innovative framework on which these capabilities may be implemented and enhanced.

One of these EIT usages enabled through the use of Intel VT is the Client Isolation and Recovery (CIR) usage model that emphasizes isolating manageability and security services in a virtual manageability appliance. IT departments benefit from this ability to isolate key services from end-user access while still maintaining the same level of flexibility and performance for end-user services. Additionally, the strategy anticipates that the manageability appliance will provide a rich environment for innovation for software vendors. The CIR usage model provides the ability to remotely manage the client PC during times when the primary operating environment is unavailable.

The other key usage models defined by EIT include Embedded PC Health, End-point Access Control, Outbreak Containment, and Agent Integrity and Assurance. The capabilities of these models are enhanced by the presence of Intel VT via isolation of the execution environments required by the IT manager compared to those exposed to the end user. In this paper we discuss how Intel VT enables a virtualized environment for a host of provisioning manageability and diagnostic applications for the IT professional.

INTRODUCTION

Intel® microprocessors and chipsets that support Intel Virtualization Technology (VT) make it feasible to run multiple operating systems (OSs) concurrently [1]. This enables the execution of multiple distinct protected execution environments that run in parallel. One such environment, the services or manageability partition, provides an isolated, controlled, and protected environment to support Embedded IT (EIT) on the platform.

EIT is Intel's strategy of embedding capabilities on the platform that enhance the overall manageability, security, and maintainability of the platform. The usages that define EIT in the business or office environment create a compelling value proposition for the use of virtualization technology on the platform. The challenges faced in the home computing environment present an opportunity to explore some of the key differentiators between the business and home computing environments.

The Intel® Lightweight Virtual Machine Monitor (LVMM) is a Virtual Machine Monitor (VMM) that partitions a client platform into two execution environments, using Intel VT, known as VT-x [2]. An execution environment is referred to as a virtual machine (VM) or a partition. One partition is the main user partition, and it can run a shrink-wrapped OS such as Windows XP*. The second partition is a services partition that runs a headless OS in an isolated execution environment. The user partition owns all the devices on the platform except for the network interface controllers. The latter are owned by the services partition, providing an ability to monitor and/or filter network traffic. Management applications that run in the services partition provide a remote console the ability to administer the client system.

In this paper we first present an overview of EIT and the capabilities that are enabled through the use of Intel VT. Next, we discuss the implications of using EIT in the home environment and follow that by an explanation of the VMM solution that we implemented for client virtualization. Finally, we conclude with a discussion of the implication of EIT on performance in the mobile environment.

EIT IN THE OFFICE

Enterprise IT departments are being asked to secure and manage an increasingly heterogeneous enterprise computing environment with less and less resources. IT departments face the need to satisfy multiple end-user client usage models and support requirements. Additionally, the IT manager faces a substantial increase in attacks to mission-critical applications and services with *for hire* attacks becoming more prevalent. As the enterprise increases in size, the scalability of existing manageability solutions is becoming a serious issue. Manageability solutions that require human intervention to discover, diagnose, and remediate system problems cannot scale to meet the requirements of large enterprise computing [4, 5]. One solution to these problems is to rely on the client platform's capability to secure, discover, diagnose, and remediate itself. In order for this to occur, manageability and security features need to be "embedded" into the client platform.

EIT Usage Models

Based on the issues IT departments face in managing their assets we came up with a set of usages that provide the capabilities required to address these issues. In this section we describe these usages and discuss how they address the challenges.

Client Isolation and Recovery

Among the challenges IT departments face today is the need to satisfy multiple end-user client usage models and support requirements. In response to these greatly varied requirements, the end user may even be granted "Administrator" rights on the client PC to install the custom software and hardware required to perform a specific job. Unfortunately, in this scenario, the end user leverages this access to install additional, non-IT validated software and hardware or disable IT security services. This results in unstable and unsecured PC configurations threatening the overall enterprise environment. Even though this additional un-validated software and hardware causes problems, end users still expect IT to support them when the client PC services and data stored on the PC become unreliable and unavailable, regardless of what the end-user Service Level of Agreement stipulates.

IT departments benefit from the ability to isolate key manageability and security services from end-user access while still maintaining the same level of flexibility and performance for end-user services. The Digital Office Embedded IT platform strategy emphasizes isolating manageability and security services to a virtual manageability appliance based on Intel VT via the CIR usage model. Additionally, the strategy anticipates that the manageability appliance will provide a rich environment for manageability and security vendors to innovate their product offerings. The CIR usage model provides the ability to remotely manage the client PC during times when the primary operating environment is unavailable. IT needs this "out-of-band" management capability in the client PCs to enable support when the end user most desires it.

The user of a CIR-enabled platform is a corporate user. The user is aware of the primary operating environment referred to as the User OS (the User Partition). IT management software runs isolated from the user's OS in its own appliance-like virtual Service Partition. In fact, the end user has no knowledge or awareness that the virtual Service Partition exists. Within the Service Partition a Service OS is used to provide an environment for IT manageability services to do the following:

- Disable malicious code or user actions.
- Prevent invalid/unsecured client configurations from adversely affecting resources on the production network.
- Patch or repair infected clients.
- Prevent situations when a worm or user deletes critical OS files.

Below are use cases of a CIR-enabled system. They are stated as problems from the customer's (either end-user or IT department) point of view.

- Virus detection and containment.
- Malicious code or the user can disable features such as Intrusion Detection, Firewall Capabilities, and Asset Management.
- Content access enforcement differs based on environment and location.
- Clients that have an invalid or unsecured configuration can adversely affect other clients on a production network.
- Infected clients cannot be patched or repaired.
- A worm or the user deleted critical OS files.

Endpoint Access Control

The Endpoint Access Control (EAC) usage, also known as Network Access Control, is a major feature of the Digital Office initiative. In the EAC, usage client access to an enterprise is contingent on the client platform being in an acceptable state. The enterprise determines the parameters of acceptability expressed in the form of an access policy. The policy is interpreted by a Policy Decision Point (PDP), which in response controls Policy Enforcement Points (PEPs) that respond by controlling access. Access controls can include any of the following:

- Unrestricted access.
- Conditional access based on traffic filtering.
- Restricted access where only specific resources are accessible.

EAC follows a methodology that can be broken down into the following general steps:

- *Collection*–monitoring, reading and storage of security measurements of the client system.
- *Reporting*–formatting collected measurements for consumption by a PDP.
- *Evaluation*–interpretation of reports and organizational policies.
- *Enforcement*–applies access control rules.
- *Remediation*–applies configuration rules designed to bring the platform into compliance.

The EIT strategy emphasizes distribution of PDP functionality to an Intel VT Service Partition through delegation. EAC policies relating to the evaluation of measurements is provisioned to a Service Partition-hosted PDP process. This process may evaluate measurements directly and forward a summary to the enterprise PDP, or measurements may be forwarded unmodified. The PDP response is interpreted by the Service Partition-hosted PDP process, and it maps the result to a format and structure that is meaningful to the client platform. See Figure 1 for the EAC architectural diagram.

The EIT strategy places a strong emphasis on locating enforcement mechanisms inside the client platform while continuing to extend control interfaces to the enterprise network. Protection of enforcement mechanisms from user applications is achieved through Intel VT to create a Service Partition. User applications and OSs function within a single User Partition. Partitioning of Host and Management environments provide isolation of EAC enforcement mechanisms and ensures that threats originating from the host environment will not defeat the goals of enterprise-controlled EAC.

The Service Partition is a *collection point* for host traffic destined for enterprise networks. Traffic filters that implement EAC enforcement policies are applied by a firewall contained in the Service Partition. Use of hardware-based filters, such as those implemented in the chipset, is under the control of the firewall process in the Service OS running in the Service Partition.

The Service Partition is an endpoint of communication for the platform. When connecting over an entrusted communication layer, a Virtual Private Network (VPN) must be constructed to establish a trustworthy connection to the enterprise network. VPN terminology broadly refers to any channel security protocol that provides data integrity or data confidentiality. A VPN therefore can be constructed at any layer in a network protocol stack. The client side of the VPN that is used for EAC originates within the Service Partition (and not the User Partition). The keys used to authenticate the endpoint and to protect channel data are managed by the Service Partition. Use of hardware-based encryption/decryption of network traffic is controlled by a VPN management process in the Service Partition. Even if packets are encrypted/decrypted in a hardware component, the logical endpoint of communication is still the VPN process.

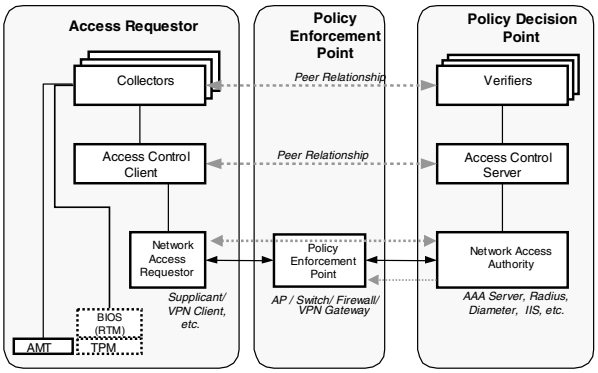

Figure 1: End point access control architecture

Outbreak Containment

Outbreak Containment (OC) provides the capability to contain the threat once an outbreak is detected. In a scenario where the client is infected, the client may be switched to a private network to enable remediation. In more serious threat scenarios, the client may be powered off to protect it from the network. In a known threat scenario, the client is updated with a patch to protect it against the outbreak.

The OC process starts when an outbreak is declared and enabled from the Management Console. The process is enabled by configuring OC filters that enable deep packet inspection for monitoring network traffic. The OC Filter Manager analyzes the collected data to assess the client health and generates a report for appropriate actions. The report can be either sent to a centralized Intrusion Prevention System (IPS), a decision-making system with a database for further analysis (Figure 2). The IPS will analyze the threat situation aggregating data from all clients. The Management Console gets the threat report from the database. If the report indicates a threat, an IT technician initiates steps to protect against the threat. In such a situation, the client is isolated from the network and a trusted out-of-band (OOB) channel is used to patch the client against the threat from network.

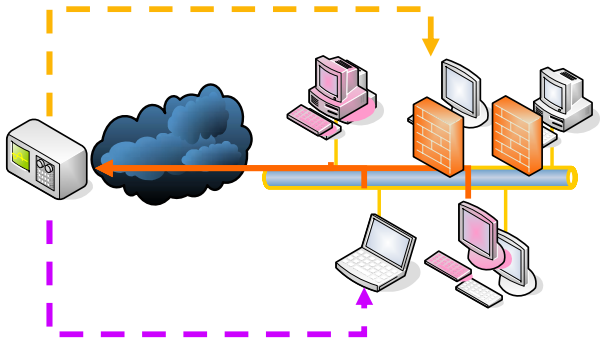

Figure 2: Outbreak containment applied at firewall

Embedded PC Health

Embedded PC Health (EPCH) usage reduces the client PC lifecycle costs by providing embedded asset management, provisioning, self-diagnostic, self-repair, and self-optimization capabilities within the Intel® platform. This OS-independent framework, based on AMT, utilizes platform-specific knowledge from Intel's processor, chipset and NIC. For details on AMT the reader is referred to [4].

This framework complements existing activities of system vendors services organizations, and manageability framework providers by adding persistent, secure, and reliable self-managing agents that support these autonomic frameworks. These programmable agents offer Independent Software Vendors and Original Equipment Manufacturers (OEMs) the ability to differentiate their offerings while benefiting from standardized capabilities and interfaces. They are accessible in an OOB mode, allowing for reconnaissance and management actions even if a PC has not yet been provisioned with an OS, or if the OS is dysfunctional.

The following is a summary of the main objectives of the EPCH:

- *Deployable*: Utilize currently deployed protocols and services in the IT environment. Minimize the need to develop and deploy new protocols and services.
- *Highly available*: Provide remote management capabilities regardless of the operational state of the PC hardware or OS.
- *OS-independence*: Provide a base set of platform management functions and interfaces regardless of the OS type or version installed on the PC.
- *Tamper-resistant*: Prevent the end user from removing or disabling the remote management service.

Security Implications

The addition of virtualization, service partition, and management processors to address security and reliability goals may seem at first counter productive due to increased overall complexity. Complexity implies greater opportunity for vulnerabilities to remain hidden and the threat of new attacks to continue.

The good news is EIT adds complexity where it is needed; it creates safer execution environments and improves the ability to detect and prevent attacks. Among these improvements is boot verification. A technique relied upon by malware is to silently install attack code into core OS files and in boot code. Each time the system boots, the malware is reinserted and

reinvoked. Anti-virus scanners are helpful, but can be spoofed by compromised OS code that lies about its existence.

In EIT systems, only code that is approved by IT or its manufacturer can be loaded. If attack code is successful in inserting itself into the system, the EIT verified boot procedure will detect the modification and apply an appropriate remediation action that can include failing to load the attack code or booting a safe-mode environment that hasn't been compromised.

Even legitimate code contains vulnerabilities that can be exploited by attackers. For example, a network driver is always subject to attacks on the networking protocols. If an attacker is successful in exploiting a vulnerability, the executable code in memory could become compromised. As systems become more reliable, they reboot less often making active attacks to memory more profitable to attackers. EIT is countering this threat by monitoring memory pages that should not change or should change in a prescribed way. Monitoring agents serve as integrity sentinels that notify the VMM whenever an invalid page access is attempted. The VMM can respond by blocking such accesses. Integrity Agents are themselves protected by a VM boundary where direct access between partitions is not allowed.

Should an attack be successful resulting in compromised EIT services, the VMM can respond by placing the platform into a more secure state. This can be achieved by alerting a management console, blocking I/O, and causing one or more VMs to cease operating. The latter is usually applied as a last resort if other corrective action fails and when a convenient time (for the user) can be identified. Automated and semi-automated dismantling of execution environments is analogous to boot verification; the core principle being that the system is always able to operate securely.

A fundamental tenet of EIT security mechanisms is the ability to create isolated execution environments that are less susceptible to attack. Intel VT and LaGrande Technology (LT) are instrumental in creating such environments. LaGrande Technology can be used to create a trusted environment even when most other parts of the system become compromised including memory, disk, and I/O. From this vantage point, it is possible to construct an environment from any remaining uncompromised components. By incorporating remediation capabilities into each primitive environment, actions can be taken that are most appropriate to the severity of the attack or failure [3].

EIT IN THE HOME ENVIRONMENT

Home IT, EIT for the home space, shares many common traits with the office EIT challenges and solutions. Like office EIT users, home users are experiencing an increasing number of attacks and spyware that degrade their experience or compromise their personal information. Home PC support organizations also spend millions of dollars, and users spend hours on the phone to diagnose and correct issues with their platforms. In many cases, both the time and monetary costs of support can be reduced using the same basic architecture as the office EIT solution, which allows the support organization to connect to the platform and diagnose problem even if the primary operating environment is unavailable. They both share a common VMM infrastructure, network isolation, and application model. The major differences are found in the connectivity model, privacy requirements, and absence of AMT.

The connectivity model differs from the enterprise model in that the managed platform is always located across the open Internet from the service provider and is highly likely to be located behind a NAT firewall. The NAT firewall presents challenges that are overcome in a couple of different ways, which are described later.

Privacy requirements in the home differ considerably from the enterprise environments. In the United States, corporate IT has complete access to a platform and all of the data stored within. This means that no permission from the user is required before IT administrators are able to access and maintain the system, access stored data, etc. In the home space, it is especially important for users to maintain control over how and when their platform and data are accessed by the IT service provider. The Home IT architecture ensures that the user is always the initiator of all service provider accesses, has visibility into all actions performed by the remote administrator, and has the right to restrict access to sensitive data.

Architectural Implications

The Home IT architecture is similar to that of Office EIT. The platform is virtualized using the LVMM and split into two VMs: the User OS (UOS) and Service OS (SOS). The SOS owns all PCI-based network devices, such as integrated Ethernet adapters, and virtualizes the devices into the UOS. The UOS owns all other platform devices, such as USB controllers, video chipsets, audio chipsets, and storage devices. A description of the LVMM is provided in the Client VMM section of this paper.

The SOS contains the Home IT control applications. The control applications work with agents in the UOS to

carry out manageability activities. The control applications use heartbeats to monitor the operation of agents in the UOS. If agents are not found to be running, a control application will signal the UOS to restart them.

The control applications communicate with the IT service provider using Web services. To request services, the control applications use Simple Object Access Protocol (SOAP) messages to the service provider. The applications accept commands from the service provider via a Web Services for Management (WS-MAN) interface. Each command is authenticated and checked for proper authentication before being routed to the proper control application for processing. Commands that require data from, or actions to be carried out in the UOS, are proxied via VMM channels to agents in the UOS, which report their status back to the SOS control applications. Figure 3 describes the architecture of Home IT systems.

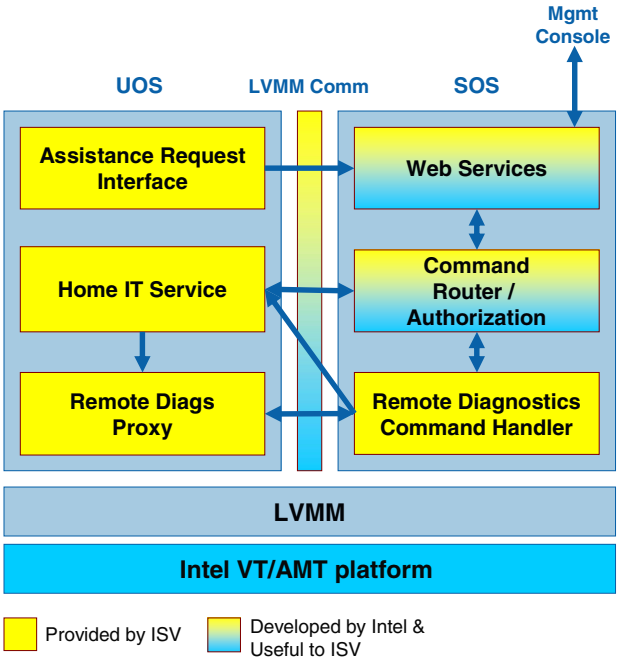

Figure 3: Home IT systems architecture

Connectivity Model

The primary Home IT connectivity model assumes that the platform will be located behind a broadband gateway router that implements a NAT firewall. In this configuration, the service provider is unable to directly connect to the SOS control applications as would be the case in the enterprise LAN configuration. The Home IT architecture includes two connection methods that can accommodate SP protocol requirements.

The first connection method is compatible with all broadband routers with NAT functionality. With this method, the service provider maintains a rendezvous server to receive incoming requests from customers. When the platform user requests assistance, the SOS control application connects to the rendezvous server and establishes a mutually authenticated Transport Layer Security (TLS) channel. Once the user's request has been sent to the service provider, the SOS processes WS-MAN commands over the channel. Service providers connect to the home platform via the rendezvous server, which proxies the commands and results between the management console and the SOS control application.

The second connection method requires the user's broadband router to support the UPnP* Forum's Internet Gateway Device (IGD) interface, but adds the flexibility of service-provider-initiated connections if necessary. For this method, the service provider maintains a registration server to receive requests from customers. When the platform user requests assistance, the SOS control application selects a random TCP port and uses the UPnP IGD interface to create a port mapping in the broadband router. The port mapping allows a connection from a specific external source to pass through the broadband router to a specific address and port on the internal network. The SOS control application then makes a Web services call to the registration server over a mutually authenticated TLS connection. The call submits the home network's WAN IP and mapped port, along with the user's problem statement and basic system configuration information, to the service provider. When a technician becomes available to assist the user, the management console obtains a waiting help request from the registration server and makes a mutually authenticated TLS connection to the user with the supplied WAN parameters. At this point, the administrator has the same capabilities as with the previous connection method. The difference is that multiple connections from the service provider to the SOS control apps can be created if necessary without involving the rendezvous server.

Privacy Considerations

In most of the world, the home environment requires a higher level of user privacy than the corporate environment. Where the corporate IT typically has complete access to a platform and its data for any reason, the home environment should afford the service provider the minimum necessary access to process user requests. This can be accomplished via a combination of policy and technical means.

Policy methods of protection require the service provider to create the Home IT software in such a way that it does not violate the user's rights. For instance, the software

should not communicate information about the user or platform to the service provider without the consent of the user. This includes user activity information and personal files. Personal files include user-created documents and anything in regions of the file system marked private by the user. Application information in the registry not directly connected to the proper configuration and operation of the program should also be out of reach of the service provider.

Technical methods of protection include encrypted or removable media for personal data, and user-specified file system filters enforced by the SOS control applications. Encrypted media is a simple solution that prevents service providers from gaining access to data by making it unintelligible, even if it is accessed by the service provider. Unfortunately, it has a number of key management complexities that could make it frustrating to the user. For instance, if the user forgets the encryption password or loses the encryption token, such as a smartcard, then the data become unreadable for the user unless there is an offline backup.

User-specified file system filters are a simple mechanism that can be enforced by the SOS control application's WS-MAN implementation. For this method, the user creates a list of file system locations that are off-limits to the Home IT agents in the UOS. The WS-MAN implementation then applies that filter to incoming commands from the service provider that either deny requests that include those file system locations, or filters the results to exclude those areas.

Authentication and Authorization

It is important to make sure that all parties and commands are properly authenticated and have the appropriate authorizations. The Home IT reference implementation uses mutually authenticated TLS and WS-security mechanisms to authenticate users. It employs access control lists in conjunction with WS-security mechanisms to enforce authorization constraints.

Authentication of communicating parties is accomplished via mutually authenticated TLS. When the SOS control applications for a platform are provisioned, they are given a unique TLS credential and one or more TLS root certificates to which the service provider components must be associated. On connection, the SOS control agent confirms that the remote party's certificate is issued by a valid root, and the service provider confirms that the SOS credential is valid and likely confirms that the user's account is in good standing. Since both parties possess a TLS credential, either one may act in the role of TLS "server" for establishing connections. If the only connection method supported by a service provider uses the rendezvous server, as described above, an alternative method of SOS authentication to the service provider exists. In this case, unilateral TLS authentication of the rendezvous server by the SOS control application can be used with a unique access token in the SOS. The SOS control application is authenticated via a challenge response exchange within the TLS tunnel. This reduces the Public Key Infrastructure (PKI) requirements of the service provider at the cost of connection flexibility.

The Home IT reference implementation uses an authorization mechanism based on WS-security with public key cryptography. For this mechanism, each administrator holds an authentication key pair used to sign WS-MAN commands. The SOS control application confirms the signature on the command and then checks an Access Control List (ACL), which includes the file system privacy filters described previously, that determines whether or not the command is permitted for that administrator. When the authorization model is fully implemented, the administrator does not have to sign each message, as the public key in the TLS client certificate shows the identity of the command issuer.

To make administration of permissions for a large number of administrators fairly easy, the Home IT authorization mechanism supports administrator groups and delegations. Administration groups allow the service provider to define administrator roles and permissions for each. The authorization mechanism then assigns one or more roles to the individual administrators. Administrators include WS-security tokens in the WS-MAN commands to assert their roles and associated permissions.

Delegation is supported in the authorization Home IT reference implementation to allow for situations where administrators may need to issue one or more commands that would normally be outside their permissions. In this case, a second administrator with the necessary permission issues a security token to the administrator handling the customer issue that grants the specific permission for a limited time. The administrator can then use that token to perform the additional action. It is up to the service provider policy to ensure that only appropriate delegations are used.

The combination of roles and delegation enable a number of useful scenarios for the service provider. For instance, the service provider can create a role for an automated triage component on the rendezvous server. The triage component would be capable of performing system inventory queries and possibly triggering routine operations such as virus or spyware scans, but would not be able to perform actions that modify the platform. Once the triage operations are complete, the case would

be passed to an administrator with the appropriate permissions to perform further diagnosis and repairs if necessary.

Another scenario involves multiple classes of administrator. In this case, the typical administrator has the permissions necessary to carry out common fixes, but not actions that are more "dangerous," including certain registry or driver modifications. If a more restricted action is necessary, the administrator must obtain a delegation of permission from a higher class of administrator, usually after the second administrator has reviewed the action for correctness.

CLIENT VIRTUAL MACHINE MONITORS

The Intel® LVMM architecture was designed with several goals in mind: maximize performance, have low complexity, maintain user experience, and provide an isolated execution environment for management applications that are always accessible and active.

High Performance

The Intel LVMM architecture was designed to maximize performance. As a result, the VMM itself virtualizes only the minimal set of devices required for allowing two distinct execution environments to execute concurrently, e.g., interrupt controllers and system timers. The LVMM allows the user partition direct access to most of the devices and therefore does not intercept I/O accesses made to those devices. This minimizes the overhead incurred by the LVMM on the user partition.

The network traffic of the user partition is handled by the services partition. The architecture depicted in Figure 4 shows that the network traffic flows through a physical NIC driver in the services partition, a bridge driver that routes the packets between the services partition network stack and the user partition network stack. In the user partition, a virtual NIC driver is responsible for sending all outgoing packets from the user partition to the bridge driver. The bridge driver forwards them to the physical NIC driver which in turn sends on the wire. Incoming packets are forwarded by the physical NIC driver to the bridge driver. The bridge driver forwards the incoming packets to the virtual NIC driver which in turn forwards them up the user partition network stack. This networking architecture provides a higher virtualization abstraction level. It performs better than a virtualization scheme that exposes a NIC device model to the user partition. In this scheme all the user partition accesses to the NIC device need to be intercepted and emulated.

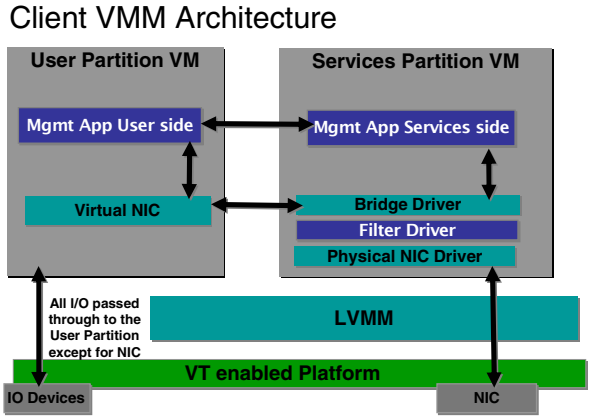

Figure 4: Client VMM architecture

The services partition has been modified to be aware of virtualization (paravirtualized). For example, the services partition interrupt handling is performed using a generic interrupt controller that is greatly simplified compared to a standard interrupt controller (e.g., Programmable Interrupt Controller a.k.a. 8259). The paravirtualization of the services partition simplifies the interaction with LVMM and saves unnecessary context transitions between them.

Unmodified User Experience

An important goal for the LVMM was to preserve the same user experience as that of a platform without the LVMM. The following are design decisions that were made in order to achieve this goal.

The design of the networking architecture guarantees that it is transparent to the end user and that the NICs are controlled by the services partition. All the functionality of the NICs controlled by the services partition, are exposed to the user partition. The virtual NIC driver in the user partition acts as a proxy for the physical NIC driver executing in the services partition. The bridge driver in the services partition provides the relay for packet data and control information between the virtual NIC driver and the physical NIC driver.

The ACPI policy decisions for configuration and power management operations of the platform are owned by the user partition. Any sleep state transition requested by the user partition is honored by the LVMM and services partition. For example, if the end user wants to transition the platform into "stand by" mode (sleep state S3) to preserve battery life, then the LVMM will eventually forward the request to the underlying platform. The user experience is preserved with respect to system power management, system thermal management, battery life, and sleep state usage models.

Isolated Execution Environment

In order to keep the partitions isolated from each other, there is a need to protect their physical memory from being tampered with by another partition. Memory accesses are performed by the CPU (e.g., any move to memory instruction) and are performed by devices through Direct Memory Access (DMA) operations. DMA allows a device with appropriate hardware to directly access system memory for data transfer without the intervention of the CPU.

The LVMM and the services partition have to be protected from memory accesses performed by user partition code. The LVMM needs to retain control over the physical memory, and thus over the processor's address-translation mechanism. We employ VT-x to prevent intentional or unintentional memory accesses from the user partition that may compromise the services partition or the LVMM.

The LVMM maintains an alternative page-table hierarchy that effectively caches translations derived from the hierarchy maintained by the OSs running in the user and services partitions. VT-x provides the necessary hooks for the LVMM to keep the alternative page-table hierarchy consistent with the OSs original page-table hierarchy. Such a hook is the trap on CR3 change. CR3 points to the base of the page-table hierarchy. Each time the OS switches to a different page-table hierarchy (i.e., changes the CR3 value), then the LVMM gets notified and switches to an alternative page-table hierarchy that matches the new OS page-table hierarchy. Since the LVMM controls the actual page tables, it can prevent a situation in which one partition has access to another partition's or the LVMM's physical memory. The LVMM prevents the existence of virtual to physical translations that map physical pages that do not belong to the partition.

The LVMM and the services partition have to be protected also from DMA bus mastering devices mapped to the user partition. These DMA-capable devices can access the entire system memory and can intentionally or unintentionally access (read/write) memory pages hosting the LVMM and services partition code and data structures. Such accesses could compromise IT secrets or render the platform useless by memory corruption. We employ Intel VT for Directed I/O (VT-d) to prevent such DMA-based attacks.

VT-d allows two views of the system memory: Guest Physical Address (GPA) and Host Physical Address (HPA). The LVMM keeps the HPA view which is the same as the system physical address space. The user and services partitions are provided their respective GPA views. The LVMM maintains shadow page tables to translate GPA to HPA for accesses from the CPU. Similarly, using VT-d DMA remapping engines and corresponding translation tables, the LVMM maintains GPA-to-HPA mapping for all DMA-capable I/O devices. Figure 5 illustrates this usage model.

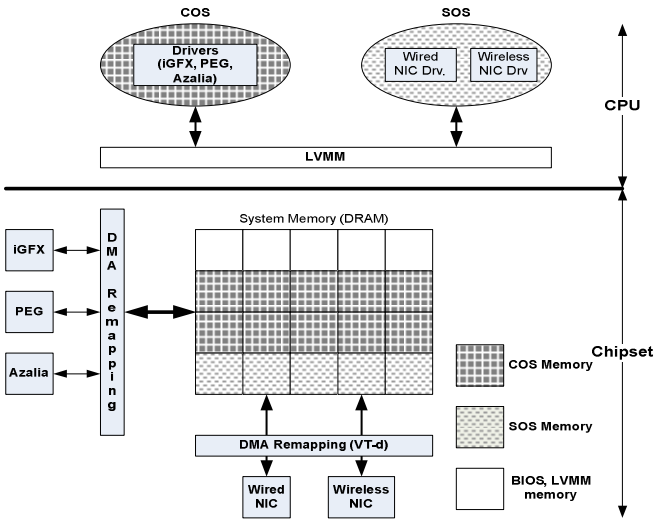

Figure 5: VT-d usage model in the client VMM

The mapping is performed as follows:

- All services partition memory pages are added to one domain such that only DMA devices mapped to services partition (NICs) can access these pages.
- All remaining pages (except LVMM and BIOS reserved) are added to the user partition domain, and all devices except those mapped to services partition can access these pages (e.g., iGFX, PCI/PCIe add-on cards etc.).
- The LVMM and BIOS reserved regions are protected from DMA accesses by virtue of being absent from the VT-d translation page tables.

The aforementioned device-to-domain mapping has the following benefits:

- I/O devices mapped to one domain can't access the memory of another domain. For example PCI/PCIe add-on cards in user partitions can't access the LVMM or the services partition.
- Device drivers in the services and user partitions run without any changes to comprehend GPA-to-HPA mapping. This translation is transparently performed by VT-d hardware when the device issues an I/O request using GPA.

If a device misbehaves by trying to access an address outside of the mapped domain, the VT-d hardware generates a fault. This fault is captured by LVMM and is indicated to the services partition. An optional management application in the services partition can

process these faults by taking appropriate actions such as displaying an error message or initiating a platform reboot, depending on the severity of the fault.

Always Accessible and Active
Management applications in the services partition are guaranteed connectivity with the external network allowing the platform to be managed even when the user partition has been isolated. The NICs are controlled by the services partition, and any action that the user partition attempts to make that can compromise the connectivity of the management applications is blocked. For example, if an action on the user partition disables the NIC, then it will get an indication that the NIC is disabled, although the real NIC remains enabled for use by the management applications.

This allows the services partition to be always accessible and reachable by a remote management console, so that management actions can be initiated.

Moreover, VT-x allows the services partition to run in parallel to the user partition. This means that the services partition is always active, and any diagnostics it runs can always be made available.

DISCUSSION

Intel LVMM vs. General-Purpose VMM

The LVMM is a custom VMM that was tailored for specific EIT usage models for the enterprise. General-purpose VMMs allow the user to create VMs that all run on the same platform. Examples of such general-purpose VMMs are VMware GSX* and Microsoft's Virtual PC*.

The main differences between the LVMM and a general-purpose VMM are as follows:

- *Number of partitions*: With a general-purpose VMM, the user can create and run multiple user partitions. In the case of the LVMM, only one user partition and a services partition are supported.
- *Partition configuration*: With a general-purpose VMM, the user can configure the execution environment of the partition. For example, the user can determine how much physical memory is allocated for the partition, and which OS to install on it. In the case of the LVMM, the execution environment of the user partition is the same as before the LVMM was installed on the system.
- *Performance*: It can be expected that with a general-purpose VMM, the performance overheads will be higher than with the LVMM. This is due to the fact that devices might not be directly assigned to the partition, and the VMM might need to emulate some devices (for use by multiple partitions).
- *User experience*: With a general-purpose VMM, the user experience is not preserved as it is with the LVMM. For example, if the end user wants to transition the platform into "stand by" mode (sleep state S3) to preserve battery life, then the LVMM will eventually forward the request to the underlying platform. A general-purpose VMM does not allow the user to directly control the platform. The power management actions are contained within the partition itself.

Mobile Performance Implications
The Intel LVMM architecture minimizes the impact on mobile battery life because, except for the network interface, the user partition owns all of the devices on the platform. A native system is a PC not virtualized, and running a single OS. Power tests executed in Intel's lab show the LVMM battery life within 95% of a native system when running local applications on the mobile machine. The LVMM achieves this by allowing the UOS to dictate the power management state of the platform.

The user partition runs an OS that complies with the Advanced Configuration and Power Interface (ACPI) specification. When the UOS requests the system to go to a certain S-state, the following occurs:

1. The LVMM intercepts the command.
2. The LVLMM sends the appropriate command back to the UOS so the UOS thinks the request has been executed.
3. The LVMM checks the SOS to see if there is any activity occurring.
    a. If yes, then the LVMM delays forwarding the S-state command to the platform.
    b. If no, then the S-state is forwarded to the platform, and the platform is placed in a lower power management state.
    c. If the UOS "wakes up" before the S-state command is executed at the platform level, then the LVMM will not change the platform power setting.

The SOS partition to support CIR, EAC, Embedded PC Health, OC, Security, and EIT in the home are all valuable usage models; however, when executed on a mobile machine, these usage models increase the number of interrupts on the system and impact battery life. Worse case, a mobile system running LVMM can use 60% of the battery life compared to a native system. In

this scenario intensive network traffic is occurring non-stop over a few hours.

Intel's LVMM architects and developers are working to improve the battery life by minimizing polling and interrupts, and optimizing memory footprint and caching. The mobile architecture team has also been educating ISVs on how mobile power management works and what they can do to ensure their software runs more efficiently to achieve better battery life.

CONCLUSION

As the usage models for virtualizations evolve, and platforms are enhanced with more capabilities, the LVMM may extend its capabilities in the following directions:

- *Multiple services partitions*–The current services partition suits the EIT manageability requirements. Other services partitions may offer other usage models such as VOIP.
- *Standard VMM API*–As more VMMs emerge in the market, there will be a need to standardize the API provided by the VMMs, so that various applications will be able to run on different VMMs.

Intel VT capabilities embedded in Intel's microprocessors and chipsets has enabled new capabilities in client systems (both desktop and mobile). In this paper we discussed how Intel VT working in concert with AMT and LT enables the design of novel solutions for embedding IT capabilities on the client platform.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of a number of people in Digital Office and Digital Home and Software Solutions Groups who were instrumental in the realization of the work described in this paper. In particular the following people have contributed to various parts of this work: Eugene Yarmosh, Ioan Scumpu, Yasser Rasheed, Abdul Bailey, Alok Kumar, Ved Shanbhogue, Romil Shah, Dhanu Agnihotri, Uttara Korad, Rao Pitla, and Shanyu Zhao.

REFERENCES

[1] Rich Uhlig, et al., "Intel Virtualization Technology," *IEEE Computer*, May 2005, pp. 48–56.

[2] Intel Corp., "Intel Virtualization Technology Specification for IA-32 Architecture," at www.intel.com/technology/vt.

[3] Trusted Computing Group, "TCG specification architecture overview," at https://www.trustedcomputinggroup.org/*

[4] Intel IT, "Reducing Enterprise Management Costs with Intel® Active Management Technology," at ftp://download.intel.com/it/digital-office/active-management-technology.pdf

[5] D. Busch, G. Bryant, B. Sayles, T. Swinford, "The Digital Office: Cross-Platform Embedded IT for Manageability, Security, and Connectivity," *Technology@Intel Magazine*, September 2004.

AUTHORS' BIOGRAPHIES

Mahendra Ramachandran is a staff architect in Intel's Digital Enterprise Group. He is responsible for driving the Embedded IT architecture definition for the Digital Office Platforms Division. Mahendra has been at Intel for eight years focusing on architecting manageability solutions for both client and server platforms. He is involved in several working groups of the DMTF. Mahendra earned his Ph.D. degree in Computer Science from Ohio State University in Columbus, Ohio. His e-mail is mahendra.a.ramachandran at intel.com.

Ned Smith is a senior security architect in Intel's Digital Enterprise Group. He is responsible for architecting security solutions for Digital Office Platforms. Ned is the co-chair of the Infrastructure Working Group in the Trusted Computing Group. He earned his M.S. degree in Computer Science from Portland State University, Portland Oregon and his B.S. degree in Computer Science from Brigham Young University, Provo, Utah. His e-mail is ned.smith at intel.com.

Matthew Wood is a security architect within the Digital Home Group, and he has extensive experience implementing authorization systems, cryptographic middleware, firewalls, and platform security features in hardware. His e-mail is matthew.d.wood at intel.com.

Sharad Garg is an architect and engineering manager within the Digital Home Group. His research interests include parallel programming, distributed file systems and storage, and manageability. His e-mail is sharad.garg at intel.com.

Jim Stanley is a senior software engineer within the Digital Home Group, and he has extensive networking, systems, and middleware programming and embedded systems experience. His e-mail is jim.stanley at intel.com.

Eswar Eduri is a software architect in SSG/IPSD. His technical interests are data communication protocols and networks, operating systems, and real-time embedded systems. Eswar has a Masters of Technology in

Electrical Engineering from IIT-Chennai, India. His e-mail is eswar.m.eduri at intel.com.

Rinat Rappoport is a software architect in SSG/IPSD. Her technical interests are processor architecture, operating systems, and real-time embedded systems. Rinat has a Masters of Science degree in Computer Science from Technion, Israel. Her e-mail is rinat.rappoport at intel.com.

Arie Chobotaro is a software engineer in SSG/IPSD. His technical interests are processor architecture, operating systems, and real-time embedded systems. Arie has a Bachelor of Science degree in Computer Science from Technion, Israel. His e-mail is arie.chobotaro at intel.com.

Carl Klotz, Jr. manages the Digital Office Platform and Solutions Development Group and he is an architect of the digital office Embedded IT appliance. He and his team have defined the solution architecture for Intel's digital office Embedded IT program. Carl is knowledgeable in various client manageability and security standards and various Intel platform technologies such as Intel Virtualization Technology, Intel Active Management Technology and LaGrande technology. Carl joined Intel in 1995 and for the past 11 years has focused his efforts on evolving Intel's platforms to be increasingly more manageable and secure. In his spare time he enjoys snowboarding, running, and memorizing long strings of random numbers. His e-mail is carl.klotz at intel.com.

Lori Janz is a program manager within the Mobile Platforms Group. She is responsible for managing the software program management team that leads all of Mobile's software platform programs as well as Mobile initiatives and technologies. Her e-mail is lori.a.janz at intel.com

[Δ] Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor.

[+] Intel® Active Management Technology requires the computer to have additional hardware and software, connection with a power source, and a network connection. Check with your PC manufacturer for details.

For further information visit:

developer.intel.com/technology/itj/index.htm